Recent Progress in Computational Science at NERSC

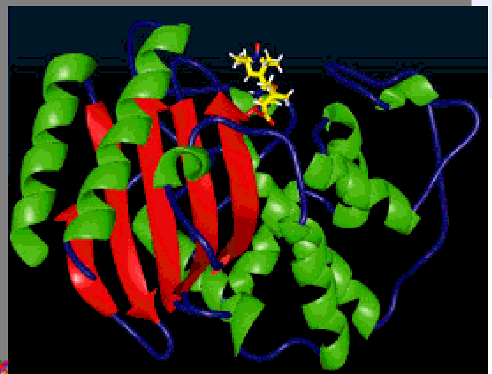
Horst D. Simon

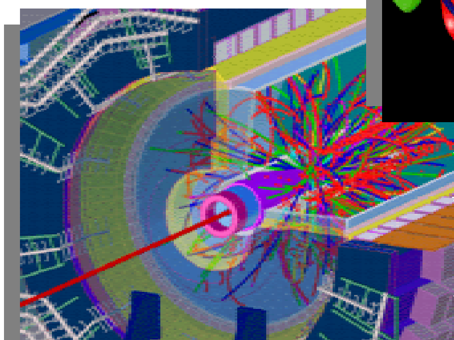
Director, NERSC Center Division, LBNL
October 3, 2003
http://www.nersc.gov/~simon

Berkeley Lab

Founded in 1931 by E.O. Lawrence on the Berkeley Campus; Moved to Current Site in 1940

National Energy Research Scientific Computing Center

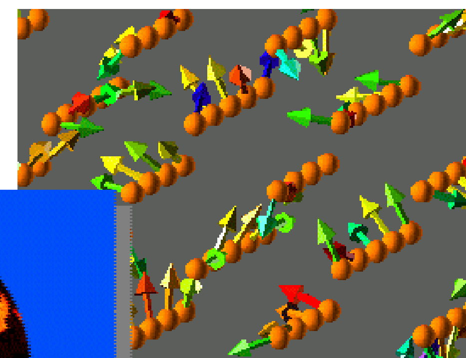


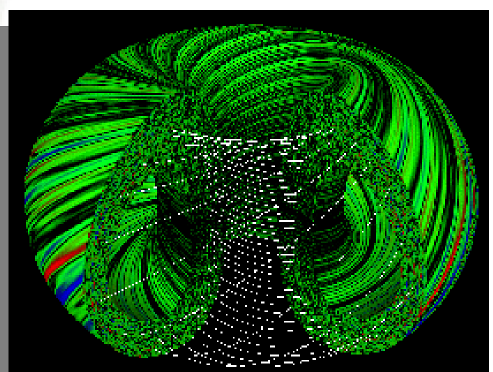


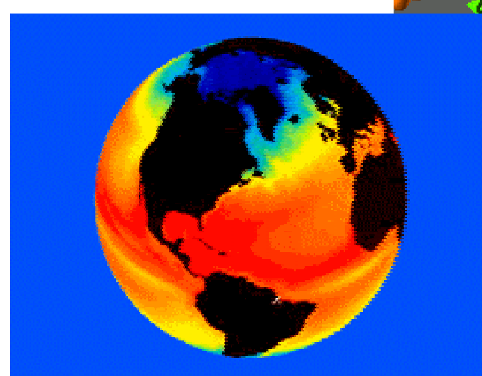
- Serves all disciplines of the DOE Office of Science

- ~2000 Users in ~400 projects

- Focus on large-scale computing







Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC Center Overview

- Funded by DOE, annual budget $28M, about 65 staff
- Supports open, unclassified, basic research
- Located in the hills next to University of California, Berkeley campus
- close collaborations between university and NERSC in computer science and computational science
- close collaboration with about 125 scientists in the Computational Research Division at LBNL

Components of the Next-Generation NERSC

Outline

- High End Systems
- Comprehensive Scientific Support – science results at NERSC
- Scientific Challenge Teams – SciDAC
- Unified Science Environment – grids

NERSC Systems

Upgraded NERSC 3E Characteristics

The upgraded NERSC 3E system has

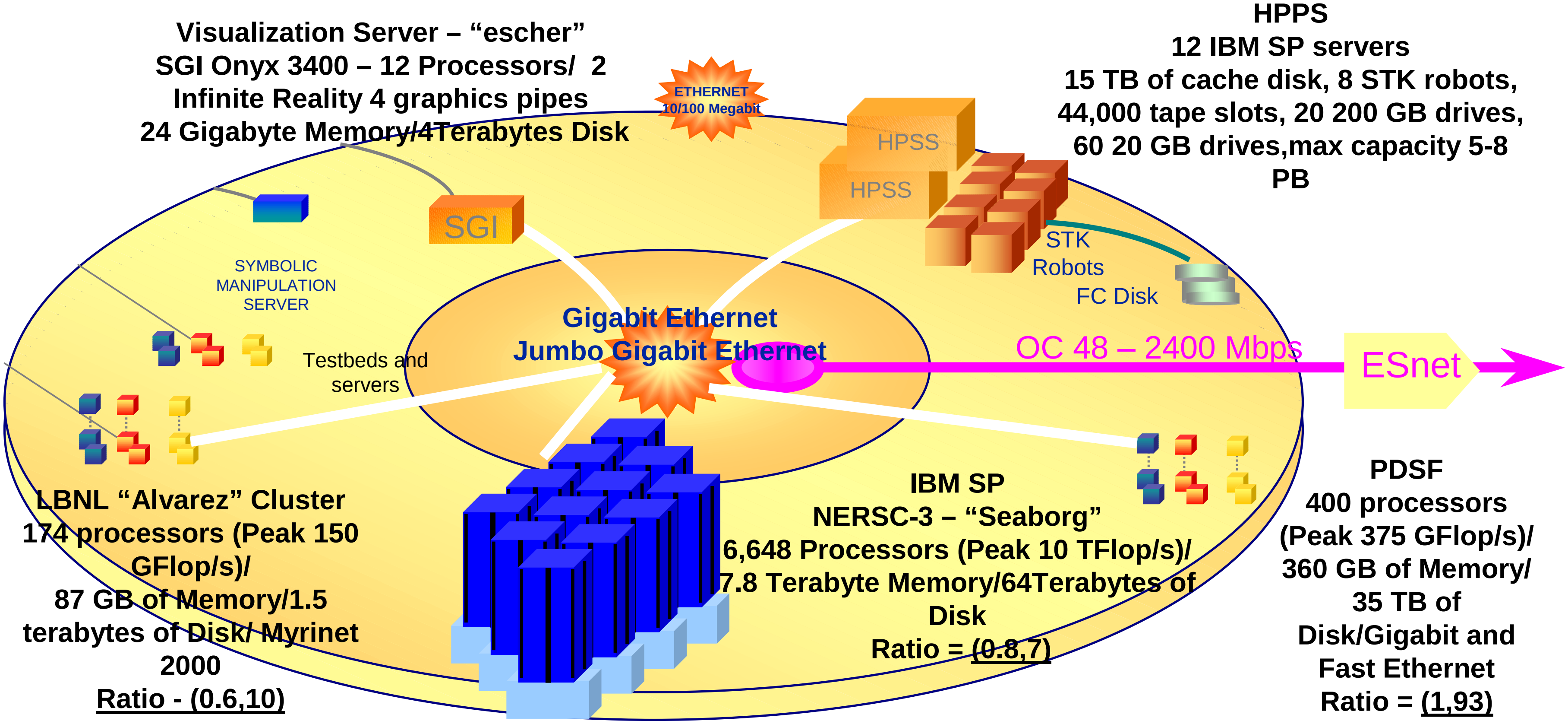
- 416 16-way Power 3+ nodes with each CPU at 1.5 Gflop/s
  - 380 for computation
- 6,656 CPUs – 6,080 for computation
- Total Peak Performance of 10 Teraflop/s
- Total Aggregate Memory is 7.8 TB
- Total GPFS disk will be 44 TB
  - Local system disk is an additional 15 TB
- Combined SSP-2 is greater than 1.238 Tflop/s
- NERSC 3E is in full production as of March 1,2003
  - nodes arrived in the first two weeks of November
  - Acceptance end of December 2002
  - 30-day availability test near completed Feb. 2003

Office of Science
U.S. DEPARTMENT OF ENERGY

TOP500 List of Most Powerful Computers

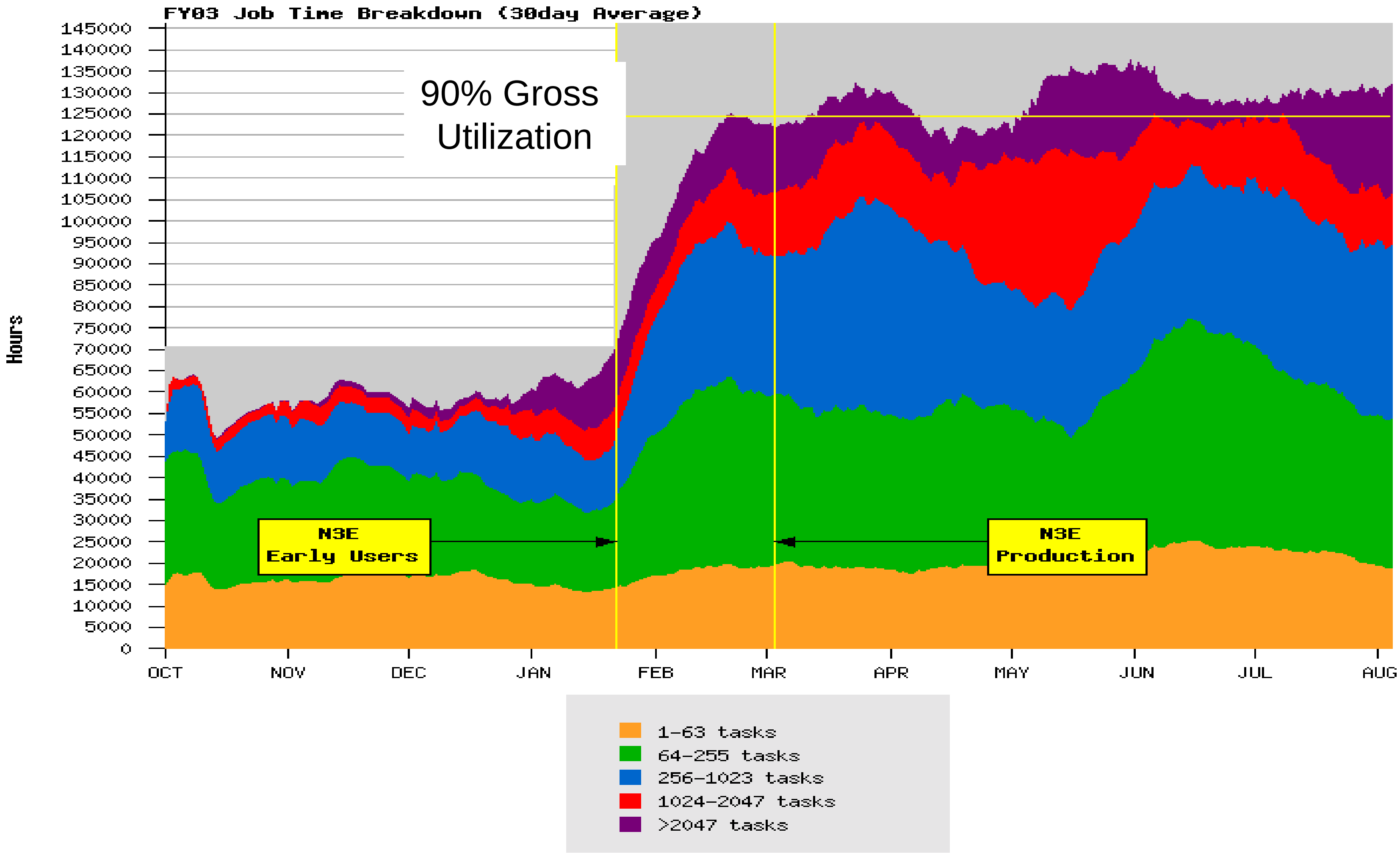
Rank	Manufacturer	Computer	R_{max} [TF/s]	Installation Site	Country	Year	Area of Installation	# Proc
1	NEC	Earth-Simulator	35.86	Earth Simulator Center	Japan	2002	Research	5120
2	HP	ASCI Q, AlphaServer SC	13.88	Los Alamos National Laboratory	USA	2002	Research	8192
3	Linux Networx/ Quadrics	MCR Cluster	7.63	Lawrence Livermore National Laboratory	USA	2002	Research	2304
4	IBM	ASCI White SP Power3	7.3	Lawrence Livermore National Laboratory	USA	2000	Research	8192
5	IBM	Seaborg SP Power 3	7.3	NERSC Lawrence Berkeley Nat. Lab.	USA	2002	Research	6656
6	IBM/Quadrics	xSeries Cluster Xeon 2.4 GHz	6.59	Lawrence Livermore National Laboratory	USA	2003	Research	1920
7	Fujitsu	PRIMEPOWER HPC2500	5.41	National Aerospace Laboratory of Japan	Japan	2002	Research	2304
8	HP	rx2600 Itanium2 Cluster Qadrics	4.88	Pacific Northwest National Laboratory	USA	2003	Research	1536
9	HP	AlphaServer SC ES45 1 GHz	4.46	Pittsburgh Supercomputing Center	USA	2001	Academic	3016
10	HP	AlphaServer SC ES45 1 GHz	3.98	Commissariat a l'Energie Atomique (CEA)	France	2001	Research	2560

NERSC System Architecture

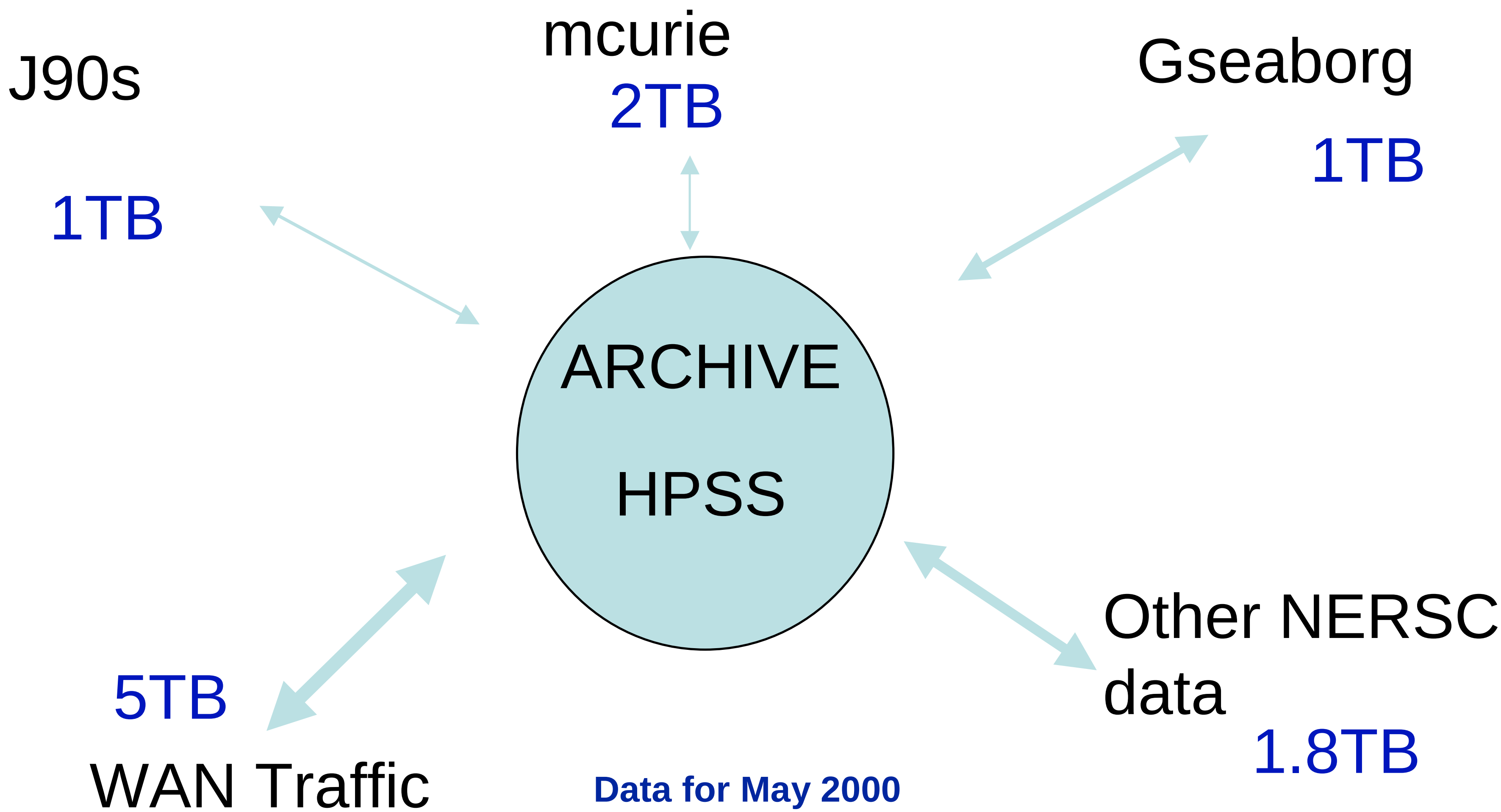


Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)

Large Job Sizes Run Regularly

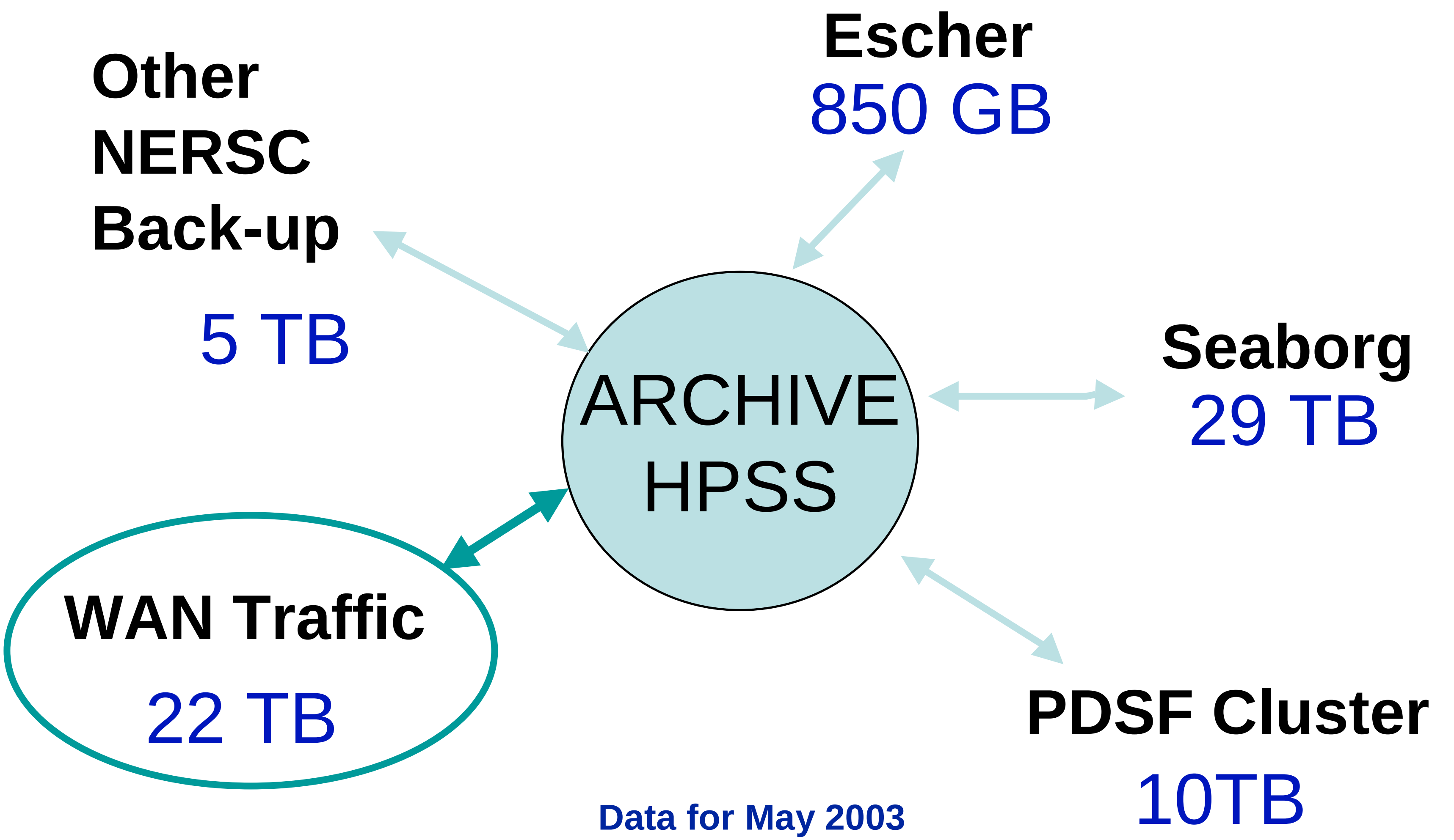


HPSS I/O Activity
May 2000 – Total = 10.8 TB



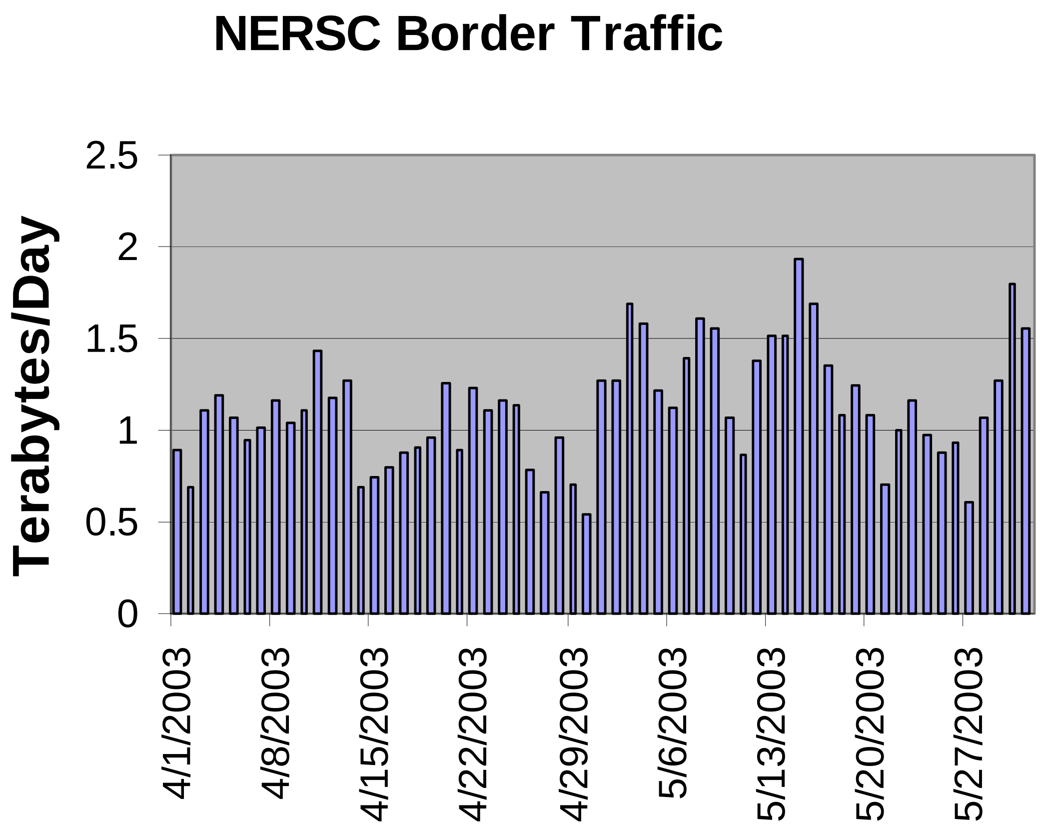
Data for May 2000

Monthly I/O Activity to Storage by Platform = 57 TB



Data for May 2003

NERSC is a Net Importer of Traffic

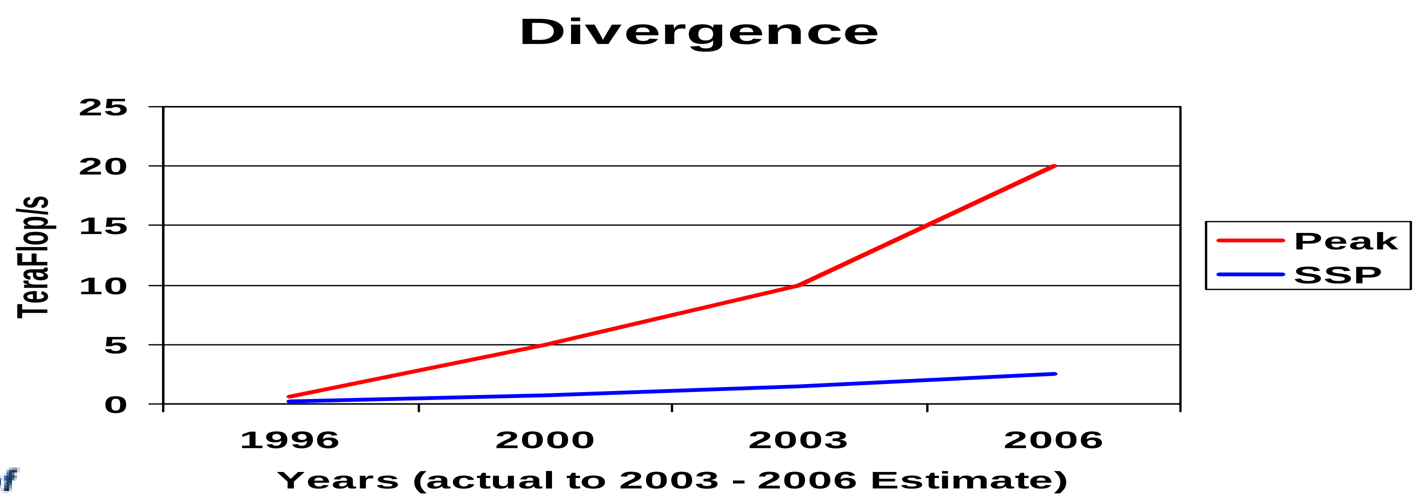


- Traffic across the NERSC border:
  - April 2003 - 29.5 TB
  - May 2003 - 39.4 TB
- NERSC traffic accounts for approximately 20% of total ESNet traffic
- 76% of the NERSC traffic is inbound

NERSC and Blue Planet

The Divergence Problem

- **The requirements of high performance computing for science and engineering and the requirements of the commercial market are diverging.**
- **The commercial-clusters-of-COTS-SMPs approach is no longer sufficient to provide the highest level of performance**
  - **Lack of memory bandwidth**
  - **High interconnect latency**
  - **Lack of interconnect bandwidth**
  - **Lack of high performance parallel I/O**
  - **High cost of ownership for large scale systems**



The Divergence Problem

- In response, NERSC, ANL, IBM developed a Science Driven Computer Architecture proposal.
  - Included a new architecture co-defined with IBM called Blue Planet
    - "Creating Science-Driven Computer Architecture: A New Path to Scientific Leadership"
  - Expanding this process with other vendors

Blue Planet

- "*Blue Planet*" is a "science driven" design process to develop systems that are simultaneously more effective for science and sustainable and cost effective for vendors.
  - White Paper uses IBM as an example of what can be done with this process
    - Can be applied to a number of vendors
- Blue Planet is a new concept for a sustainable computer architecture more effective for science and engineering applications
  - A specific implementation leveraging the IBM roadmap that better balances scientific processing needs and the commercial viability
  - Described as a "ultrascale" scale system on the order of the Earth Simulator

http://www.nersc.gov/news/blueplanet.html

and

http://www.nersc.gov/news/ArchDevProposal.5.01.pdf

Approach

- Engage the vendor community with a new approach to leveraging their major R&D/product roadmaps to create new architectures that are much more effective for science
  - Study applications critical to DOE Office of Science and others. For example:
    - Material Science, Combustion simulation and adaptive methods, Computational astrophysics, Nanoscience (new drugs and also new microchip technologies), Biochemical and Biosciences (protein folding/interactions), Climate modeling, High Energy Physics (particle accelerators and astrophysics), Multi-grid Eigen solvers and LA methods
  - Identify key bottlenecks found in these critical applications
  - Outline a high level approach to address the challenges
  - Follow-up meetings for detailed drill down by the vendor experts, computer scientists and application scientists at NERSC
  - Iterate on proposed solution

Needs Based on Scientific Applications

	AMR	Coupled Climate	Astrophysics		Nanoscience	
			MADCAP	Cactus	FLAPW	LSMS
Sensitive to global bisection	X	X	X		X	
Sensitive to processor to memory latency	X	X			X	
Sensitive to network latency	X	X	X	X	X	
Sensitive to point to point communications	X	X				X
Sensitive to OS interference in frequent barriers				X	X	
Benefits from deep CPU pipelining	X	X	X	X	X	X
Benefits from Large SMP nodes	X					

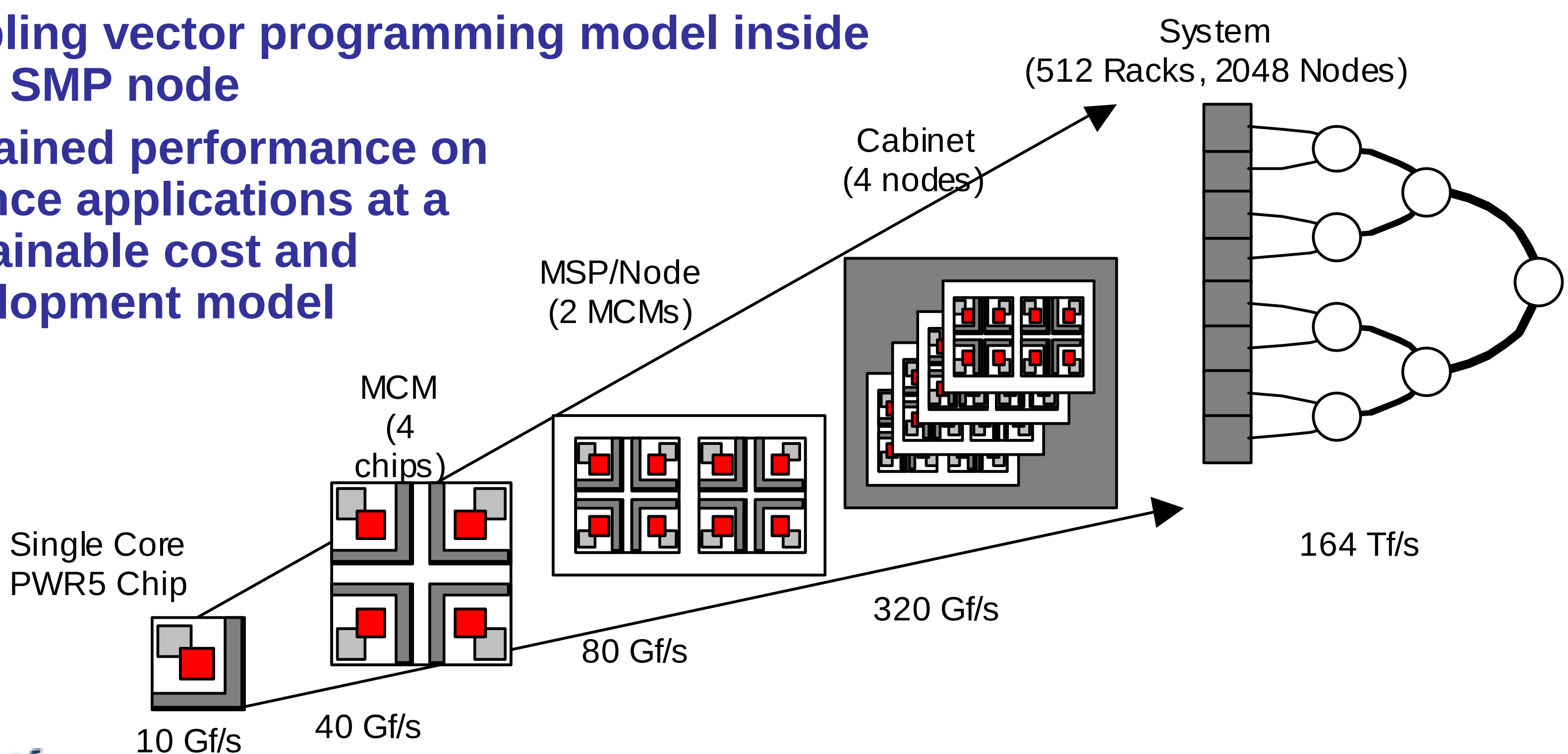
Full IBM Blue Planet System Components

- New IH++ Wide Node - 8 CPUs per node
  - POWER5 GS Processor - 2.5GHz
  - Single core MCM
- 2048 node system (8 Nodes per frame)
  - 16K processors @ 10GF per CPU = 160TF Peak
- Virtual Vector Architecture - VIVA
- Federation Switch - 3 stage topology
  - 8GB/s per server for the uni direction communication bandwidth.
- 40-50 TF Sustained on 2-3 selected applications
- 256 TB of memory = 16GB per CPU
  - May reduced to 128TB of memory if it can sustain full memory BW
- 2.5PB disk in I/O system [approximately 48 IO nodes]
- Approximately 600 Frames
  - 256 compute racks, 250 Disk racks, 160 Switch racks
  - 12,000-15,000 Sq Feet; 5-7 MWatts Power
- Scientists will focus on application optimization

Office of Science
U.S. DEPARTMENT OF ENERGY

Blue Planet: A Conceptual View

- Increasing memory bandwidth – single core
  - 8 single CPUs are matched with memory address bus limits for full memory bandwidth
- Increasing switch bandwidth – 8-way nodes
- Decreased switch latency while increasing span
- Enabling vector programming model inside each SMP node
- Sustained performance on science applications at a sustainable cost and development model



Scientific Results using NERSC

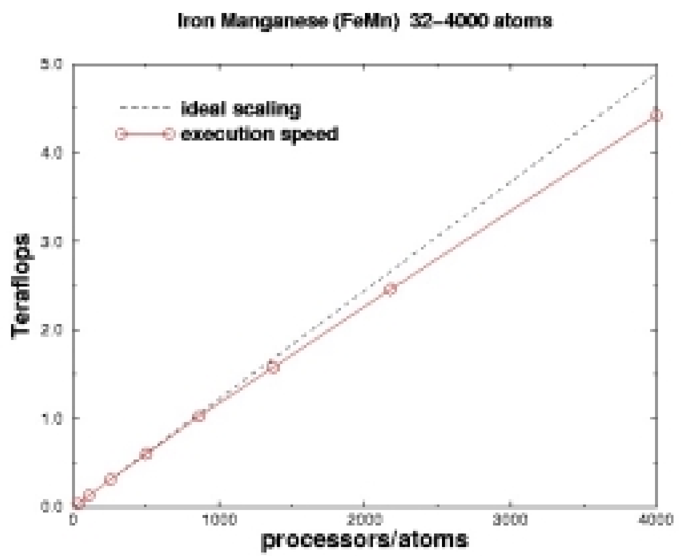
Multi-Teraflops Spin Dynamics Studies of the Magnetic Structure of FeMn and FeMn/Co Interfaces

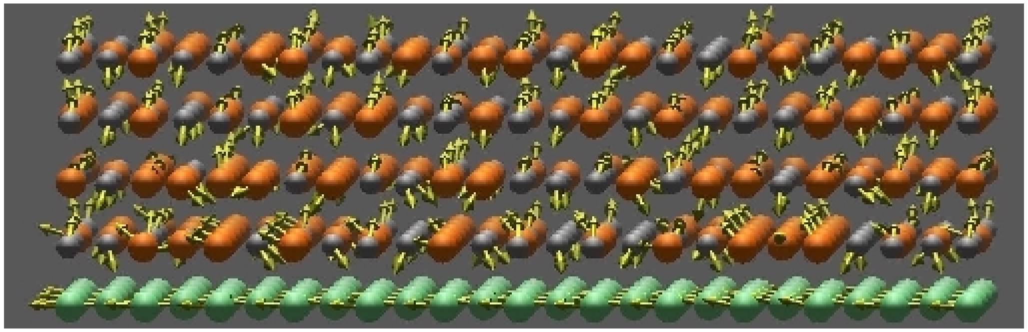
Exchange bias, which involves the use of an antiferromagnetic (AFM) layer such as FeMn to pin the orientation of the magnetic moment of a proximate ferromagnetic (FM) layer such as Co, is of fundamental importance in magnetic multilayer storage and read head devices.

A larger simulation of 4000 atoms of FeMn ran at 4.42 Teraflops on 250 nodes.

(ORNL, Univ. of Tennessee, LBNL(NERSC) and PSC)

IPDPS03 A. Canning, B. Ujfalussy, T.C. Shulthess, X.-G. Zhang, W.A. Shelton, D.M.C. Nicholson, G.M. Stocks, Y. Wang, T. Dirks





Section of an FeMn/Co (Iron Manganese/ Cobalt) interface showing the final configuration of the magnetic moments for five layers at the interface.

Shows a new magnetic structure which is different from the 3Q magnetic structure of pure FeMn.

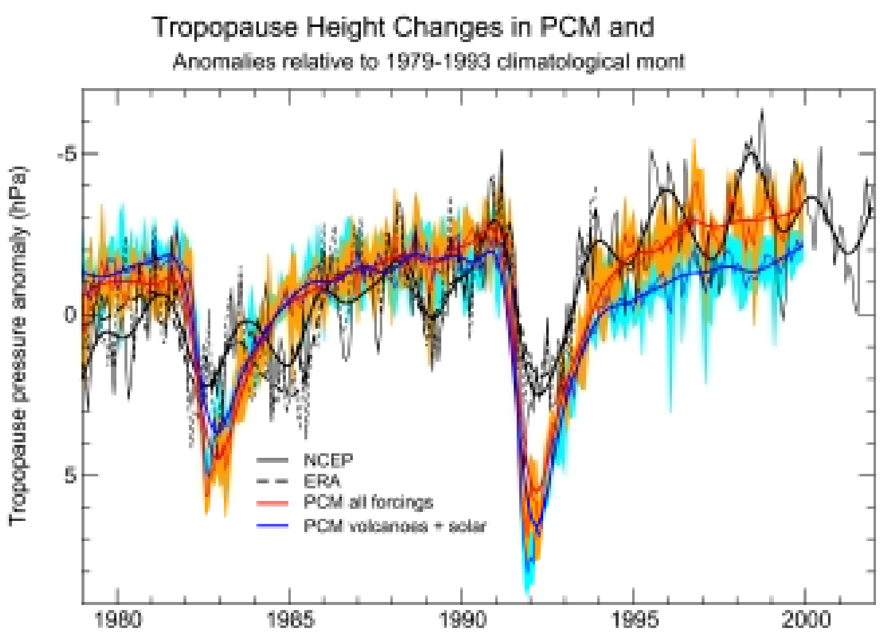
Office of Science
U.S. DEPARTMENT OF ENERGY

New Results in Climate Modeling

- Recent improvements in hardware have reduced turnaround time for the Parallel Climate Model
- This has enabled an unprecedented ensemble of numerical experiments.
  - Isolate different sources of atmospheric forcing
    - Natural (solar variability & volcanic aerosols)
    - Human (greenhouse gases, sulfate aerosols, ozone)
- Data from these integrations are freely available to the research community.
  - By far the largest and most complete climate model dataset
  - www.nersc.gov/~mwehner/gcm_data

Investigating Atmospheric Structure Changes with PCM

- The tropopause is that height demarking the troposphere and the stratosphere.
  - Below the tropopause, the temperature cools with altitude.
  - Above the tropopause, the temperature warms with altitude.
- A diagnostic that is robust to El Nino but sensitive to volcanoes.
- An indicator of the total atmospheric heat content
- Changes in natural forcings alone (blue) fail to simulate this feature of the atmosphere, but natural + anthropogenic changes (orange) do



Santer et al. Figure 1

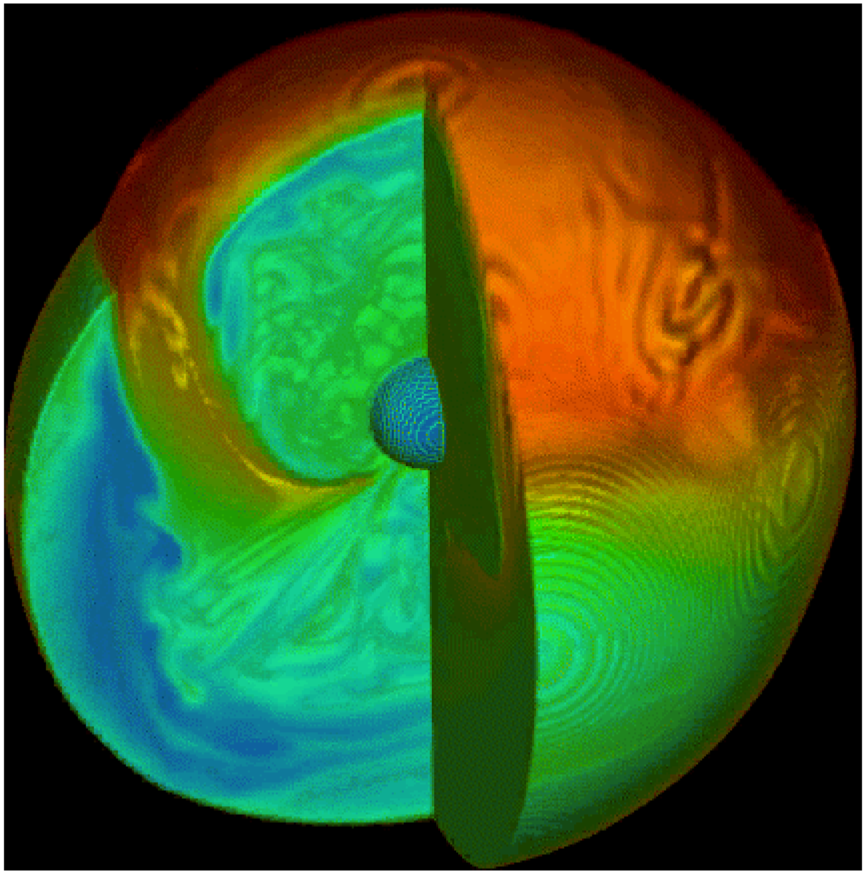
NERSC Support Efficient Science of Scale

Project	Performance (% of peak)	CPU Count
Terascale Simulations of Supernovae	35%	2048
Accelerator Science and Simulation	25%	4096
Electromagnetic Wave-Plasma Interactions	68%	2048
Quantum Chromodynamics at High Temperature	13%	1024
Cosmic Microwave Background Data Analysis (pre and post processing)	50%	2048 & 4096

Note – these are comparable to the best documented efficiencies of the science codes on the Earth Simulator, but on different codes of course.

Office of Science
U.S. DEPARTMENT OF ENERGY

Terascale Simulations of Supernovae



- PI: **Tony Mezzacappa, ORNL**
- Allocation Category**: SciDAC**
- Code**: neutrino scattering on lattices (OAK3D)**
- Kernel**: complex linear equations**
- Performance**: 537 Mflop/s per processor (35% of peak)**
- Scalability**: 1.1 Tflop/s on 2,048 processors**
- Allocation**: 565,000 MPP hours; requested and needs 1.52 million**

Office of Science
U.S. DEPARTMENT OF ENERGY

Getting the Physics out of KamLAND Data

- Solar neutrino experiments at Super-K and SNO suggested that the three flavors of neutrinos are actually different states of the same particle. If the same oscillations were found in neutrinos or anti-neutrinos from terrestrial sources, this conclusion would be confirmed.
- In January 2002 KamLAND, the world's largest anti-neutrino detector, began generating about 200 GB of data per day—too much for the network connection to Tohuku University—so the data was stored on LTO-format tapes.

- After six months, U.S. scientists running experiments at KamLAND had 48 TB of data on 800 tapes but no way to access them.

Getting the Physics out of KamLAND Data

- KamLAND tapes were shipped from Japan to Oakland, where NERSC staff had just developed software for LTO interface with HPSS.
- Researchers used PDSF to analyze KamLAND data.
- KamLAND results were reported in September 2002, confirming Super-K and SNO.
- How NERSC did it:
  - Large-scale resources
  - Operational flexibility
  - Client-oriented services

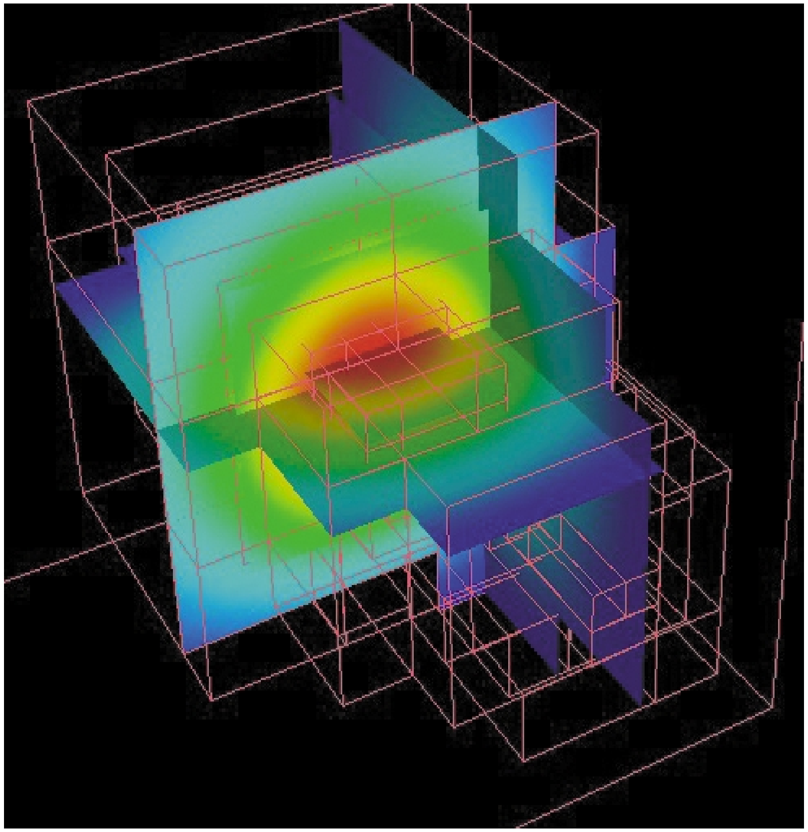
Office of Science
U.S. DEPARTMENT OF ENERGY

SciDAC

--

Bringing all resources together

High Performance Computing Research Department (HPCRD)



Juan Meza, Department Head

Groups:

- Applied Numerical Algorithms
- Center for Computational Sciences and Engineering
- Future Technologies
- Imaging and Informatics
- Scientific Computing
- Scientific Data Management
- Visualization

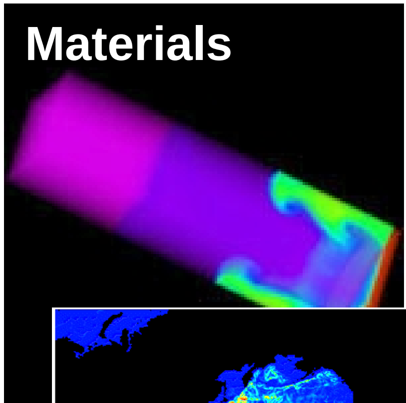
Total Staff: 108

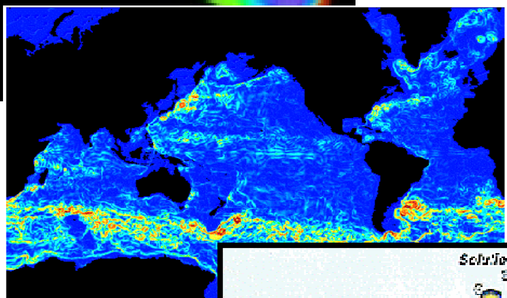
The High Performance Computing Research Department conducts research and development in mathematical modeling, algorithmic design, software implementation, and system architectures, and evaluates new and promising technologies.

Scientific Discovery Through Advanced Computing

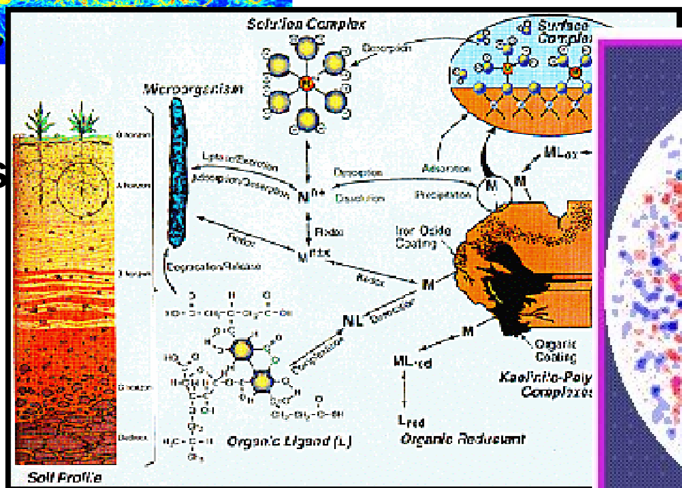
DOE Science Programs Need Dramatic Advances in Simulation Capabilities To Meet Their Mission Goals

Materials



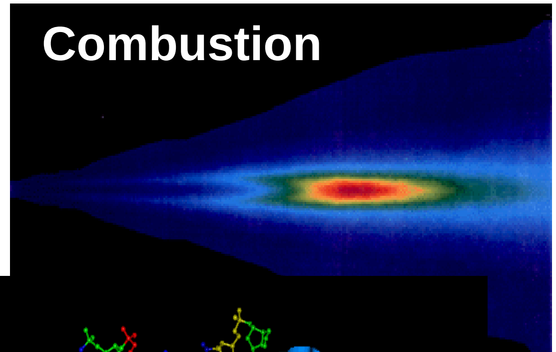


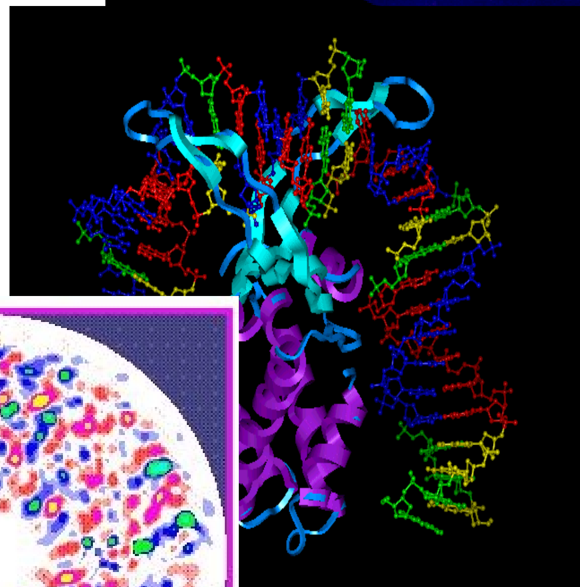
Global Systems



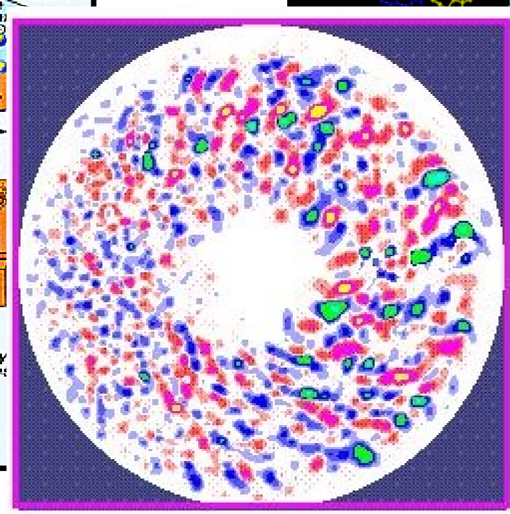
Subsurface Transport

Combustion





Health Effects, Bioremediation



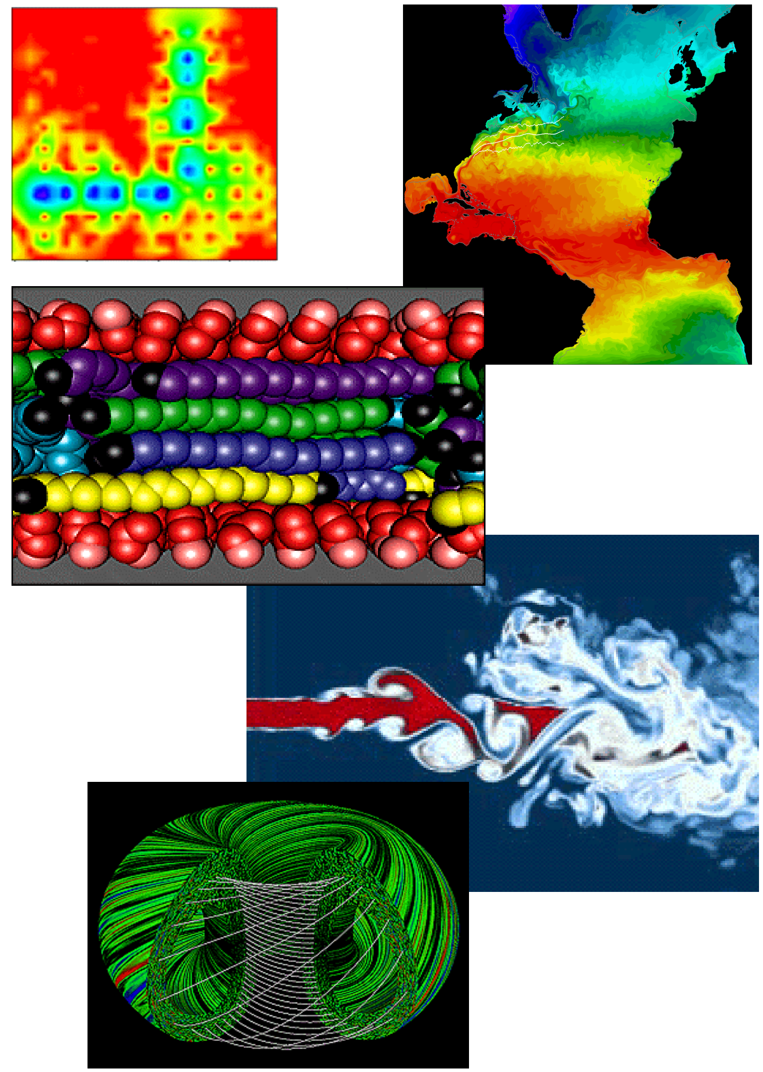
Fusion Energy

Introduction – What is SciDAC?

- SciDAC is a pilot program for a "new way of doing science"
- first Federal program to support and enable "CSE" and (terascale) computational modeling and simulation as the third pillar of science (relevant to the DOE mission)
- spans the entire Office of Science (ASCR, BES, BER, FES, HENP)
- involves all DOE labs and many universities
- builds on 50 years of DOE leadership in computation and mathematical software (EISPACK, LINPACK, LAPACK, BLAS, etc.)

SciDAC

- **Harness the power of terascale super-computers for scientific discovery:**
  - Form multidisciplinary teams of computer scientists, mathematicians, and researchers from other disciplines to develop a new generation of scientific simulation codes.
  - Create new software tools and mathematical modeling techniques to support these teams.
  - Provide computing & networking resources.



Addressing the Performance Gap through Software

Peak performance is skyrocketing

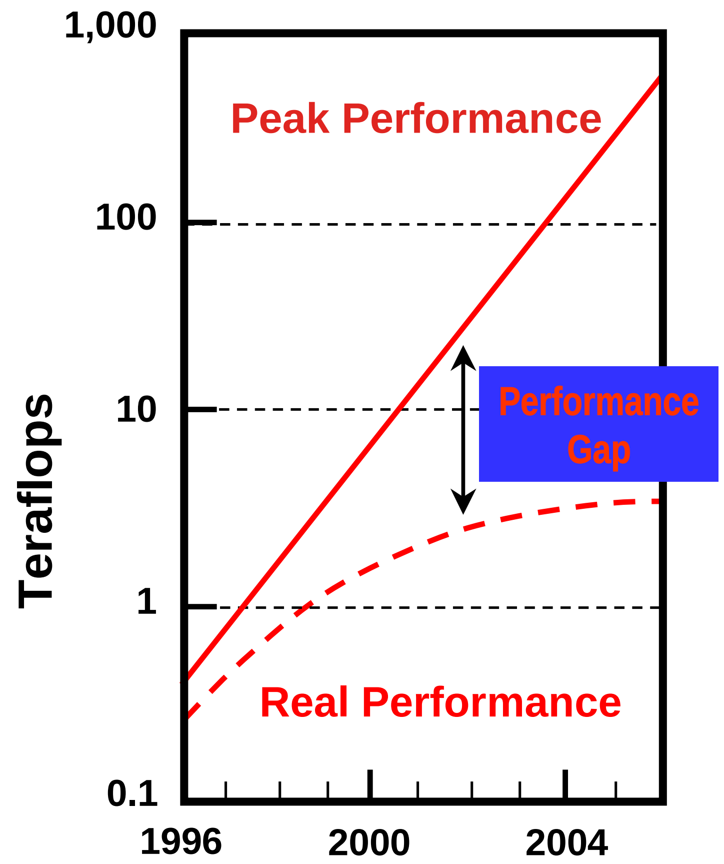
- In 1990s, peak performance increased 100x; in 2000s, it will increase 1000x

But ...

- Efficiency for many science applications declined from 40-50% on the vector supercomputers of 1990s to as little as 5-10% on parallel supercomputers of today

Need research on ...

- Mathematical methods and algorithms that achieve high performance on a single processor and scale to thousands of processors
- More efficient programming models for massively parallel supercomputers



SciDAC Focus on Software

Applications

Global Climate
Computational Chemistry
Fusion
- Magnetic Reconnection
- Wave-Plasma Interactions
- Atomic Physics for Edge Region

High Energy/Nuclear Physics
- Accelerator Design
- QCD
- Supernova Research
- Neutrino-Driven Supernovae
  and their Nucleosynthesis
- Particle Physics Data Grid

Computer Science

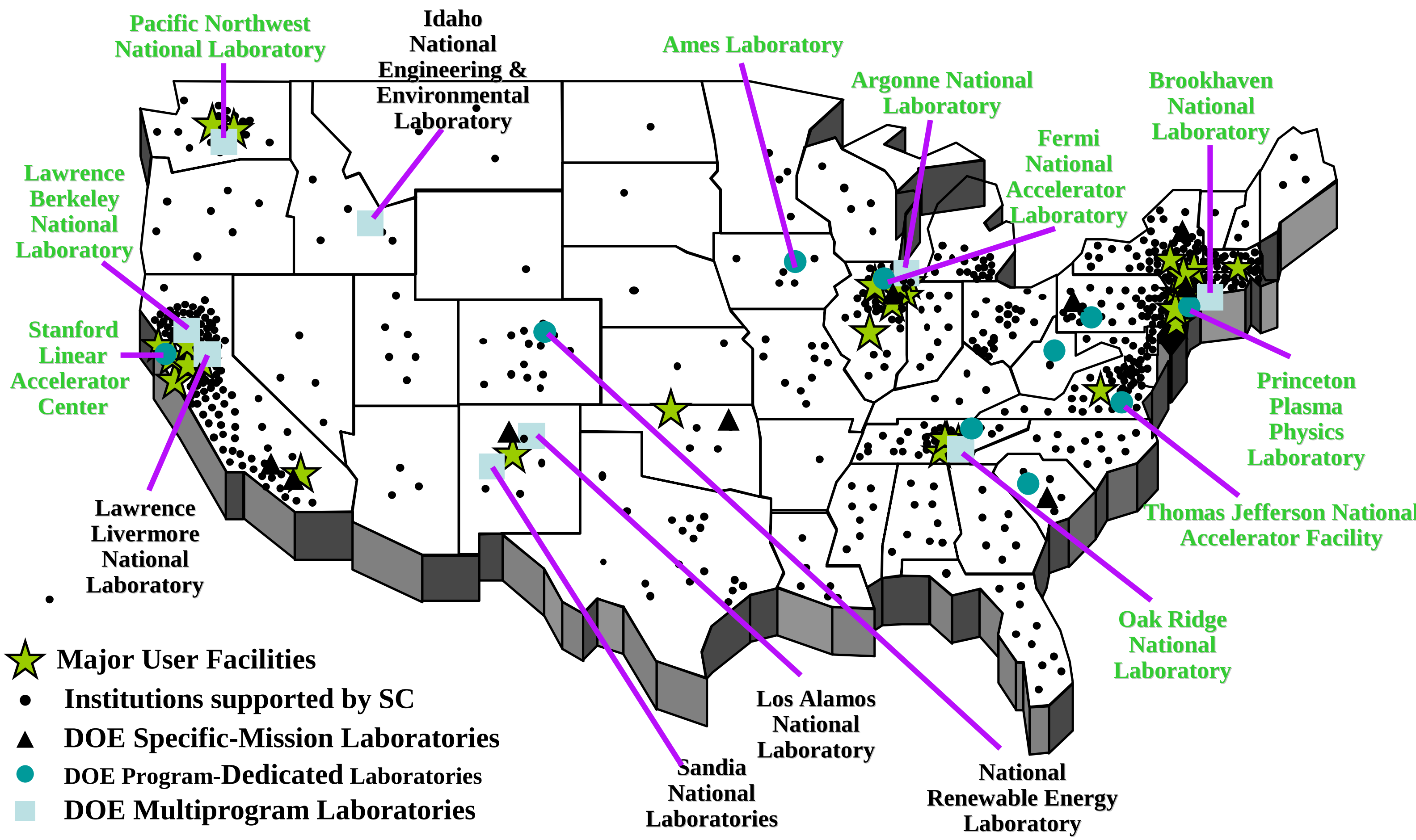
Scalable System Software
Common Component Architecture
Performance Science and Engineering
Scientific Data Management

Mathematics

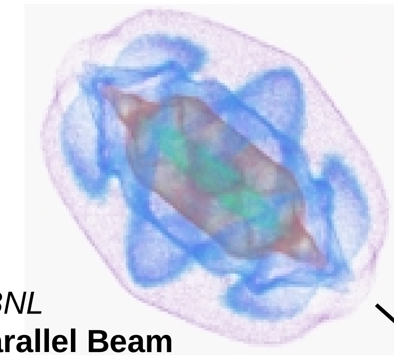
PDE Solvers/Libraries
Structured Grids/AMR
Unstructured Grids

7 Integrated Software Infrastructure Centers (ISICs) were established in FY01 (3 in Berkeley)

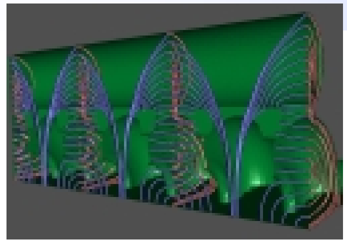
Science in the 21st Century is Distributed!



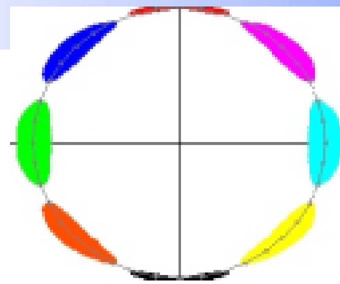
Typical SciDAC Application Project: Advanced Computing for Twenty-First Century Accelerator Science and Technology



LBNL
Parallel Beam Dynamics Simulation



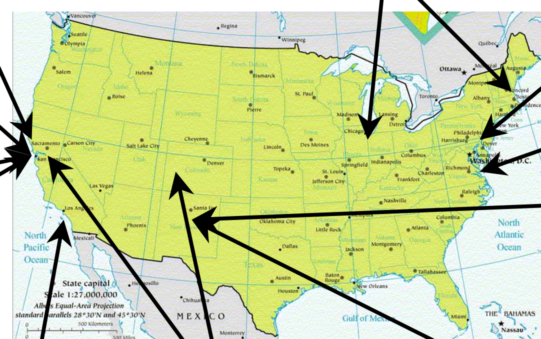
UC Davis
Particle & Mesh Visualization



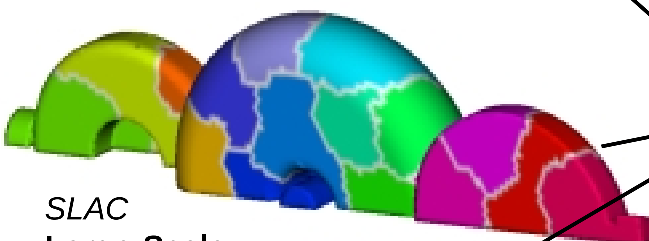
FNAL, BNL
High Intensity Beams in Circular Machines

$$M=e^{:f_2:}\; e^{:f_3:}\; e^{:f_4:}\ldots$$
$$N=A^{-1}\, M\, A$$

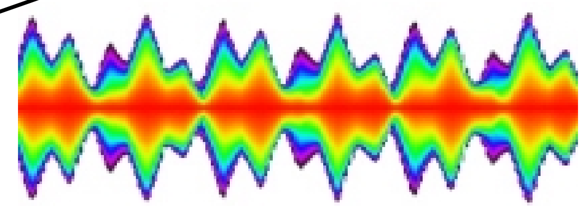
U. Maryland
Lie Methods in Accelerator Physics



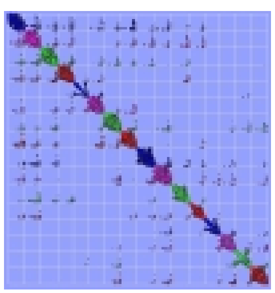
Jefferson Lab.
Coherent Synchrotron Radiation Modeling



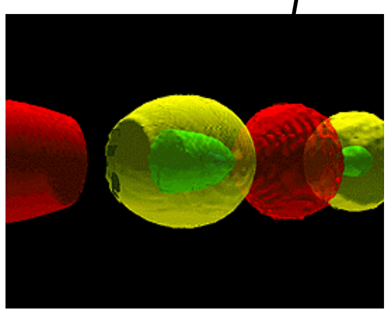
SLAC
Large-Scale Electromagnetic Modeling

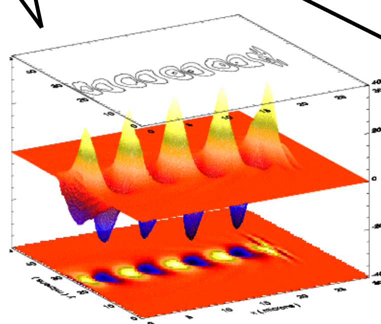


LANL
High Intensity Linacs, Computer Model Evaluation

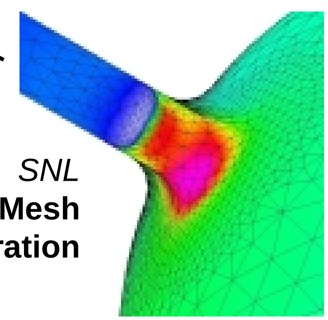


Stanford, NERSC
Parallel Linear Solvers & Eigensolvers





UCLA, USC, UCB, Tech-X, U. Colorado
Plasma-Based Accelerator Modeling

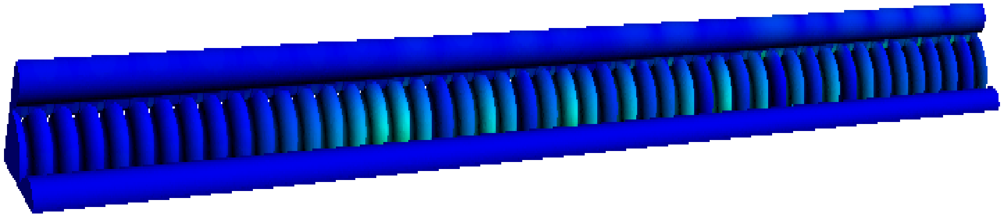


SNL
Mesh Generation

Applied Math.Contribution to Accelerator SciDAC: Large-scale Eigenvalue Calculations

- **Calculates cavity mode frequencies and field vectors.**
  - Finite element discretization of Maxwell's equations gives rise to a generalized eigenvalue problem.
  - When losses in cavities are considered, eigenvalue problems become complex (and symmetric).
  - NERSC, Stanford collaboration (PI Kwok Ko, SLAC)
    - Parry Husbands, Sherry Li, Esmond Ng, Chao Yang (NERSC/TOPS+SAPP).
    - Gene Golub, Yong Sun (Stanford/Accelerator).

Individual cells used in accelerating structure

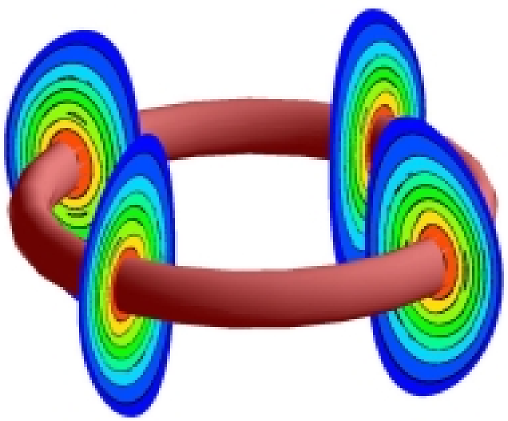


Omega3P model of a 47-cell section of the 206-cell Next Linear Collider accelerator structure

Future Applied Math. Contributions

- SuperLU:
  - Improve the interface with PARPACK.
  - Parallelize the remainder of the symbolic factorization routine in SuperLU – guaranteeing memory scalability, and making the exact shift-invert algorithm much more powerful.
  - Fill-reducing orderings of the matrix.
- Need to improve the Newton-type iteration for the correction step, as well as the Jacobi-Davidson algorithm:
  - SuperLU has its limitations: memory bottleneck.
  - Future plans include joint work (LBNL+Stanford) on the correction step.
    - Iterative solvers.
    - Preconditioning techniques.

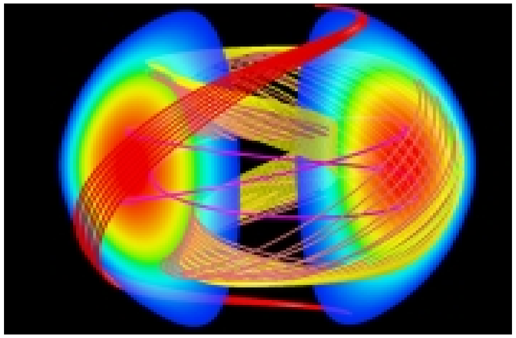
SCIDAC Collaboration Speeds Up Fusion Code By Factor of 10



- NIMROD is a parallel fusion plasma modeling code using fluid-based nonlinear macroscopic electromagnetic dynamics.
- Joint work between CEMM and TOPS led to an improvement in NIMROD execution time by a factor of 5-10 on the NERSC IBM SP.
- This would be the equivalent of 3-5 years progress in computing hardware.

- Parallel SuperLU, developed at LBNL, has been incorporated into NIMROD as an alternative linear solver.
  - Physical fields are updated separately in all but the last time advances, allowing the use of direct solvers.

    SuperLU is >100x and 64x faster on 1 and 9 processors, respectively.
  - A much larger linear system must be solved using the conjugate gradient method in the last time-advance. SuperLU is used to factor a preconditioning matrix resulting in a 10-fold improvement in speed.



http://w3.pppl.gov/CEMM

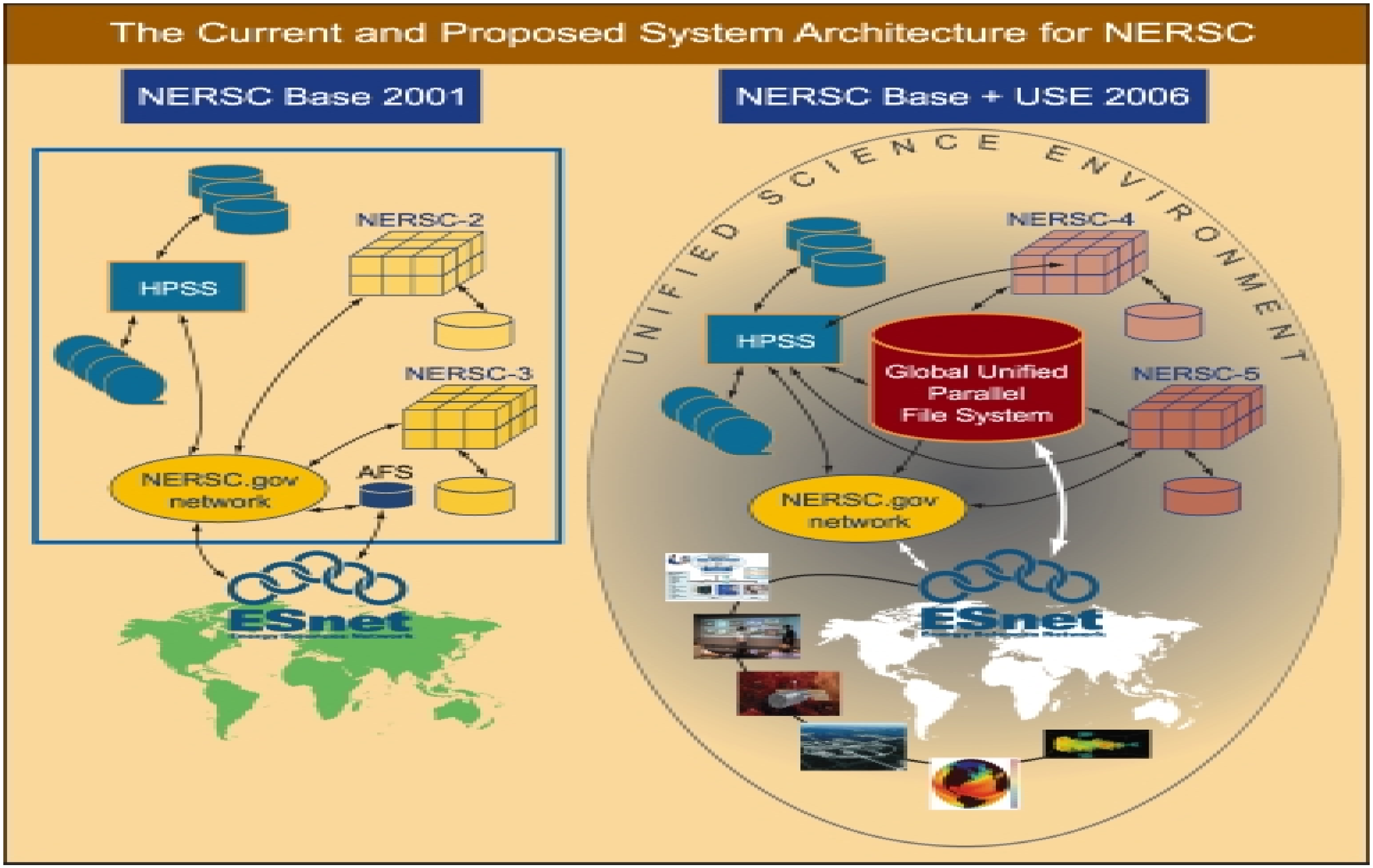
http://www.tops-scidac.org

SciDAC is first Full Implementation of Computational Science and Engineering (CSE)

- **CSE is a widely accepted label for an evolving field concerned with the science of and the engineering of systems and methodologies to solve computational problems arising throughout science and engineering**
- **CSE is characterized by**
  - Multi - disciplinary
  - Multi - institutional
  - Requiring high end resources
  - Large teams
  - Focus on community software
- **CSE is not “just programming” (and not CS)**
- **Ref: Petzold, L., *et al*., Graduate Education in CSE, *SIAM Rev*., 43(2001), 163-177**

Connecting to DOE Science Grid

NERSC Systems Will Evolve



Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC and the Grid

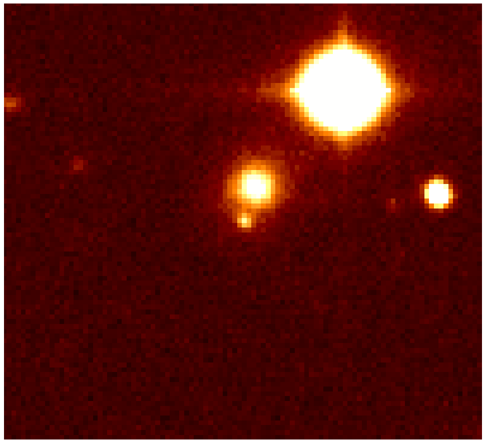
- **Multi-year plan**
  - 2002
    - **Data Grid pre-production activities**
    - **Track computational grid, collaboration, and workflow development**
    - **Established a collaborative agreement with IBM to accelerate deploying Grid Technology**
  - 2003
    - **Focus on data Grid production rollout**
    - **Pre-production compute Grid**
    - **Track collaboration and workflow development**
    - **Earth Systems Grid Prototype**
  - 2004
    - **Focus on compute Grid production rollout**
    - **Pre-production collaboration and workflow**
  - FY2005
    - **Focus on collaboration and workflow production rollout**
  - FY2006
    - **All major USE/GRID components on NERSC production systems**

Latest Activities

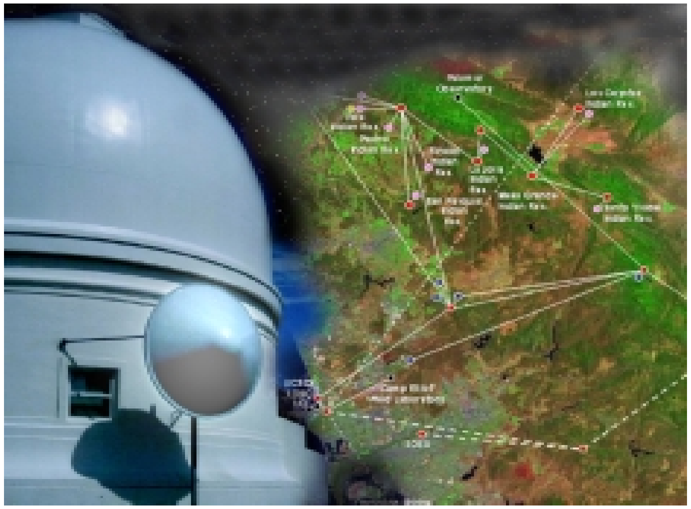
- Infrastructure is in place on all system
  - LDAP, CAs, basic globus functionality, etc.
  - Working in cooperation with IBM to test, improve and field GTK 2.2 on the IBM SP – now a few early beta users have access
  - Testing the Grid with firewalls
  - Implement Grid aware IDS features
- Production use of the Grid for Storage and PDSF
- Developed an interim solution to grid enable HPSS.
  - Now being distributed until the new GridFTP for HPSS is available
- White papers
  - Security
  - Implementation Issues
  - Vision for HPSS and the Grid

Nearby Supernova Factory

- Goal: Find and examine in detail up to 300 nearby Type Ia supernovae
  - More detailed sample against which older, distant supernovae can be compared
- Discovered 34 supernovae during first year of operation and now discovering 8-9 per month
- First year: processed 250,000 images, archived 6 TB of compressed data
- This discovery rate is made possible by:
  - high-speed data link
  - custom data pipeline software
  - NERSC's ability to store and process 50 gigabytes of data every night

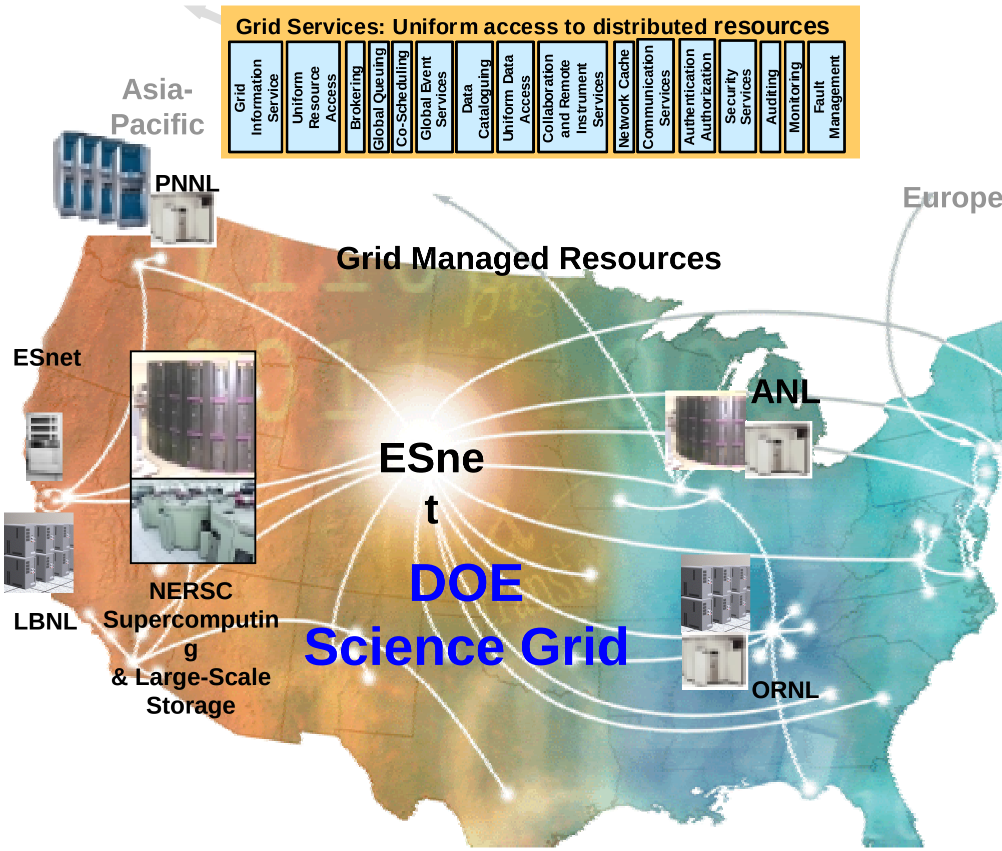


Nearby Supernova Factory



- Every night, images from the Near Earth Asteroid Tracking program (NEAT) at Mount Palomar and Maui are sent to NERSC via ESnet and a special link in SDSC's High Performance Wireless Research and Education Network (HPWREN)
- Custom data pipeline software automatically archives images in NERSC's HPSS
- Image subtraction software running on PDSF sifts through billions of objects to find supernovae
- Follow-up spectrographic observations are obtained the next night and sent to NERSC and other centers for analysis
- First major discovery: First detection of hydrogen in the form of circumstellar material around a supernova

Distributed Systems Department



William Johnston,
Department Head

Deb Agarwal, Department Deputy

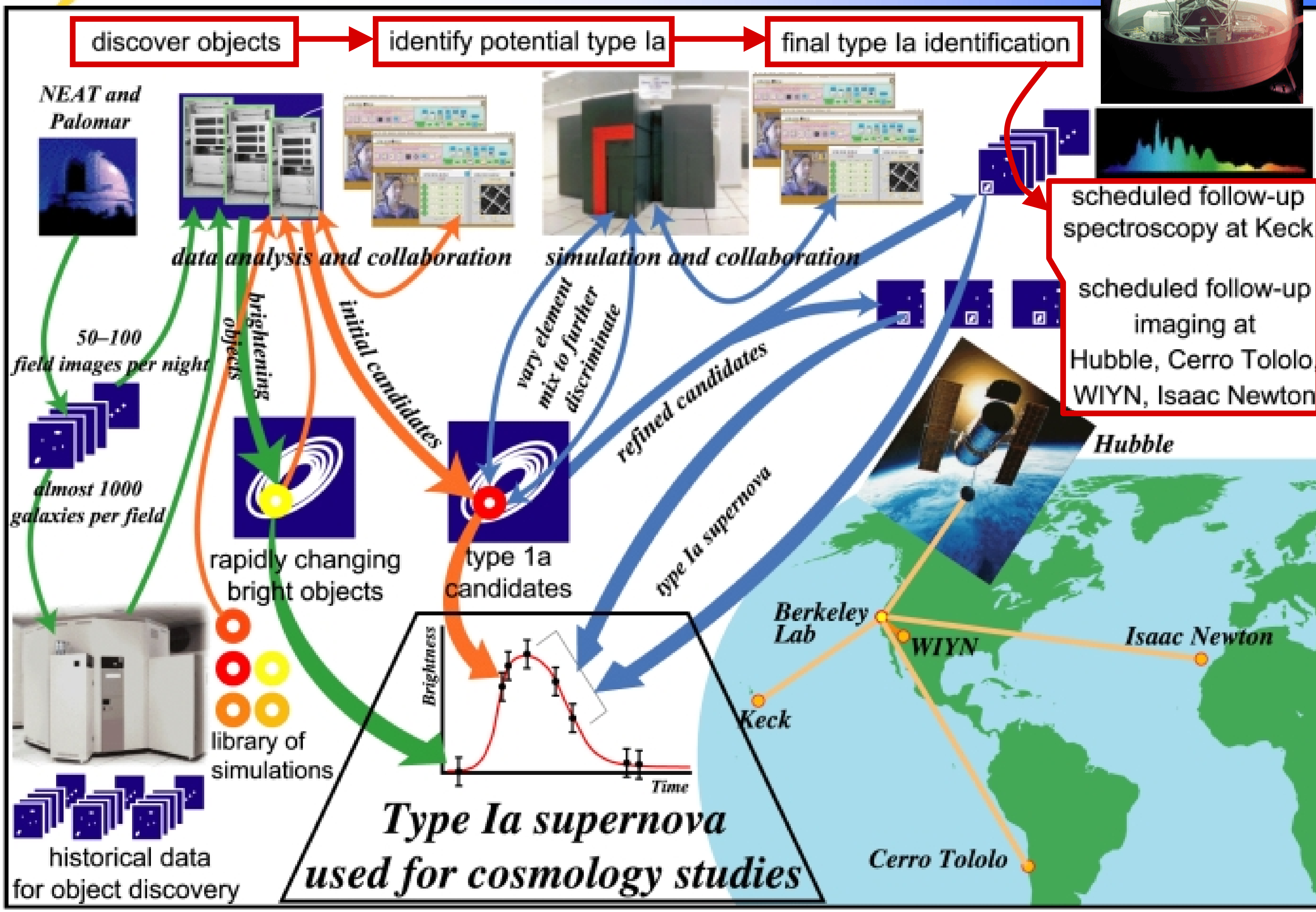
Groups:

- **Collaboration Technologies**
- **Data Intensive Distributed Computing**
- **Network Technologies**
- **Secure Grid Technologies**

Total Staff: 25

The Distributed Systems Department researches and develops software components that allow scientists to address complex and large-scale computing and data analysis problems in a distributed environment such as the DOE Science Grid.

Supernova Cosmology Depends on Cyberinfrastructure



Supernova Cosmology Project, Perlmutter, et al. (http://www.supernova.lbl.gov)

Earth Systems Grid

- A large database of PCM and CCSM results have been postprocessed and quality controlled for easy distribution to the scientific community.
  – 125 registered users for 2003
  – 24,000 SRUs of file space
- Over 80 PCM runs
  – Atmospheric monthly and daily data.
  – Oceanic monthly data.
- CCSM2.0.1 control run (years 350-999)
- See http://www.nersc.gov/projects/gcm_data/ for details or email mfwehner@lbl.gov

The Future

More resources:
No limits to growth in demand for supercomputer resources seen

Better integration:
Computational science and engineering will become recognized as discipline

Next level simulation science:
Large scale simulation environments will emerge that allow computer simulation at unprecedented scale